A theory of suffering

Consciousness, bias, AI...

and what happiness has got to do with it

...suffering is the sole origin of consciousness.

– Fyodor Dostoevsky, *Notes from Underground*

I, of whom I know nothing, I know my eyes are open because of the tears that pour from them unceasingly.

– Samuel Beckett, *The Unnameable*

'Would you like to relive the very last conscious moment in this life?' Interestingly, under this condition only a little over 28 per cent of life moments were ranked as positive, while just below 72 per cent were considered not worth reliving.…

– Thomas Metzinger[1]

If happiness is only noticeable against expectations of unhappiness – or, at least, non-happiness, we must either not require much in the way of happiness to persist with our lives, or we have psychological mechanisms for filtering from experience *awareness* of its lack. If the instants that comprise our *normal* moments we don't describe as happy ones, yet we remain unwilling to say that our lives are, by and large, unhappy ones – lives where suffering or indifference predominates, what

1."Suffering," Thomas Metzinger, in *The Return of Consciousness: A new science on old questions*, Kurt Almqvist & Anders Haag, eds. 2017, p, 221.

does this say about our reasons for being here? We have been here a long time and show no signs of quitting existence *en masse*... It seems happiness can't have much to do with it. Survival is not concerned with it. Evolution did not optimize for it. Given that we spend so much time *less* than happy, it seems we should ask what suffering is, and how it came about. In the process, we may learn what role ephemeral, positively valenced states, such as joy and happiness, play in the economy of survival.

Philosopher of mind and theoretician of consciousness Thomas Metzinger cites empirical evidence that suggests that it can't be happiness that keeps us going. If it were and we were rational, we would have given up long ago. We'll examine his theory in more detail later, but first, we remark that there are *kinds* of consciousness and that there are *kinds* of unpleasant experiences and that these are interestingly correlated. Two main kinds of consciousness are *access* and *reflective*. Two kinds of negatively valenced experiences correspond to these, respectively, and can be distinguished as *pain* and *suffering*. Pain we associate with access consciousness, and suffering with reflective consciousness. We will mostly be concerned with reflective consciousness and suffering. The suggestion is that suffering and reflective consciousness evolved together. We suffer because we are reflectively conscious and we are reflectively conscious because it has been survival, not happiness, that has driven us onward – thus far. The suggestion of a link between this kind of consciousness and the advanced form of negatively valenced state we call suffering is hardly new, though attempts to be *rigorous* about describing the relationship between the two have not, it appears, until recently, gotten much attention. Metzinger has been at the forefront of exploring the suffering/consciousness connection. He wants to sharpen the concepts involved. This is important for advancing its scientific study, but also has deep and broad ethical implications for everything biological or non-biological that behaves *as though* it were conscious.

Beyond Good and Evil, part 4, no. 146, trans. Judith Norman.

Compared to pain, the phenomenon of suffering has received far less attention by philosophers of mind, neuroscientists, and research psychologists.[2] (Within the wider humanities, however, the opposite may be true.) Pain's closer tie to neurophysiology makes it more vulnerable to empirical examination than suffering. Pain is more open to unambiguous observation, while, except in the vaguest terms, suffering has been difficult to pin down conceptually. What exactly is it? It is a negative experience most will agree.[3] But beyond that, what? Metzinger, among others, thinks that the intervention of *reflective* consciousness makes the difference between pain and suffering. Perhaps suffering and reflective consciousness, both[4] *empirically* opaque

2. Metzinger suggests this has a lot to do with a prevailing bias against bad news: *optimism bias*, a variant of *existence bias:* the most powerful bias of them all, and a major barrier to scientific progress. We evolved to survive – *for awhile* – not to do science: a consideration which must color our conclusions – however careful we are to sterilize our methods of "nonobjective" artifacts.
3. That suffering is *always* negative, however, is questionable, as we shall see below. This complicates matters. Pain, unmediated by consciousness, perhaps, is more unambiguously negative.
4. Consider the "Other Minds" Problem.

phenomena (suffering less so than consciousness because of its empirical markers), may help illuminate each other. You may ask a subject if they are suffering, and elicit a response either way, but with enough observation you may mount evidence to support a conclusion – probably because of an *affective* response on *your* part, the observer, as much as any you may infer in the subject. You are not (normally) indifferent to a judgment of suffering's presence as you may be to the color of their underwear. Asking whether they are conscious is not answered by a yes or no (or some observation as with a corresponding question about pain), but, presumably, by *any* answer at all. This suggests that cognition or an interposed judgment is required, not merely an observation. If we propose an *empirically assessable* theory of suffering, something we may study using scientific methods, something we have been less than successful proposing for the study of consciousness, *and* if suffering and *reflective* consciousness (the kind of interest here) are intimately connected, then we may be in a better position to understand the latter. This is Metzinger's program: spelling out an empirically assessable understanding of suffering, and, thereby, of consciousness.

The implications of understanding suffering and reflective consciousness and the relationship between them extend beyond human curiosity or well-being. What we learn from studying these still vague notions will have implications for animal and machine ethics.

Metzinger's theory

Are people happy with their lives? Suppose we discover that for fully two-thirds of their waking hours,[5] if asked, they report they are *not* happy. How would we measure "being unhappy"? As a first attempt at a psychometric, Metzinger proposed, in his small study, asking if subjects are willing to repeat random moments of their lives again and again forever: create "an eternal playlist" of moments. On being surveyed about their responses, most were surprised at how few of their moments were desirably repeatable: only about a third.[6] How did they react? They told stories. What would motivate them to go on with their lives, otherwise? Surely, their enthusiasm to reproduce, if not also to perdure themselves, would be affected? On being confronted with evidence that they were not happy most of their conscious lives, they quickly offered excuses to the effect that there are many *other important things*, besides whether they were happy, that explain their behavior in extending their lives and generating more like it...[7] We may conclude that either happiness, per se, is overrated as a reason to live, that it is optional, that these other reasons given are real and potent explanations for a persistent will to continue. Or, we might think these, now *informed*, explanations are tantamount to rationalizations that have proven themselves evolutionarily protective of the survival of the gene pool. These excuses evolved as psychological mechanisms to this end. Thus, they would be deeply ingrained pro-life *biases*.

---

5. Assuming that while asleep the question cannot be asked or answered meaningfully.
6. And there is reason to believe that even this figure may be an overestimation. Metzinger's study was small and consisted of mostly graduate students. Needless to say, funding to scale up such a study to a more generally representative size would be substantial – and *who* would materially profit by confirming that most people do not think their lives are that great? The odd melancholy humanist may think so, such as those quoted at the start, but it is not, nor likely ever will be, a popular move to dwell on human misery. (And, perhaps, part of what impedes consciousness research is its negative valence acquired *by association* with suffering).
7. There is a curious parallel to the short list of surmised motivations for most, if not all, ostensibly "moral" acts: habit, self-interest, fellow-feeling, and fear of consequences – *none of which* are proper moral motivations, according to Kant, due to their contingency. There are natural explanations back of all these, *all the time*, that do not involve autonomous rational agency. There can be no morality proper without the latter according to Kant: at best, only moral *behavior*: at best legality, not morality, conformity, not conviction. Would you like it if your significant other only flawlessly *acted* as though they cherished you – but didn't? Would it *matter* to you?

Pain was present from the beginning, from the first stirrings of self-preservation, but suffering never needed to have happened had we never "fallen," never become *reflectively* conscious in the first place, had, instead, we remained aware of only *this* moment and what it asked of us. Survival, then, ceased to be sufficient all by itself as it seemed it was for eons for the rest of the inhabitants of the biosphere. We hanker for something no other living thing has or needs: meaning, purpose – or, a way of escaping the need for these, but the very *need* to escape them itself only exacerbates awareness of the precarity of the condition we are in: a perennially uneasy one.

Consciousness and suffering are intimately related. Awareness of this has been around since the earliest recorded history. There is a telling passage in the *Epic of Gilgamesh*, the oldest surviving piece of literature in the Western world, a millennia or two before the earliest books of the Bible, that links sex, death, and an epistemic fall from a state of blessed ignorance that these bookends of life frame. Suffering opens eyes, raises consciousness. *Post*-awareness, nothing is ever the same again; we can no longer even accurately imagine what it was like *before* this awareness. Loss becomes subject to never being forgotten or buried by subsequent experience. A veil falls between us and the animal world with whom our affinity, once visceral, is now distanced. Thereafter, we may – but also *must* – *observe* the animal world – indeed, *all* of the world – from a lonely perch that we cannot genuinely descend from, though we may try. The verb "observe" is used advisedly. It involves a degree of abstract contemplation. Animals *see*, conscious beings may also *observe*.[8] To observe is not just to notice something because it may serve some immediate or foreseeable interest you have: that much we have in common with most animals. Rather, we mean by observation to imply a first step toward an *interrogation* of your condition that – provided one doesn't seek to numb or distract ourselves from the inquiry, or exhausted, settle for less than we might self-knowledge – is doomed to frustration. This awareness of your state, your placement in the world, the point of it, or lack thereof, is beyond painful. It is a condemnation of sorts. It is the beginning of suffering. *If* life is to carry on, it is to the benefit of those afflicted with it not to be much burdened with such a realization.

Pain and suffering

For contrast, we will mean by "pain," an experience that involves sentient capacities, but without entailing anything of their situatedness. Moreover, pain, in proportion to its severity, may quickly drown out capacities for contemplation, let alone interrogation, if these are present at all. Nietzsche suggested *extreme* pain causes *the loss* of consciousness, perforce, leaving no room for the *reflective* sort.

Suffering, on the other hand, is awareness on top of *pain*. It is to know that it is you, specifically, that are, at once, in pain *and* witness to it, that *you* have a *desire* to make the pain go away, to *know* that you have such a desire, to *recall* that there was a time when you were not in pain, and that you want to *believe* there will be again such a time. It involves the capacity to *remember, abstract, and imagine* counterfactual situations – situations that cannot be present for you *now* – if you are, in the throes of pain.

Recovering from these episodes offers the opportunity to experience more of them later. But each time the experience will either cause you to retreat therapeutically further from consciousness or erode your

8. We do not preclude the likelihood that many animals, in some degree, also "observe" in the sense we use, only that we have the clearest evidence for this in the animals *we are.* See, for instance, Peña Guzman on animal suicide.

reasons for being here. The beautiful time when you didn't need "reasons" for being here – that time before falling into consciousness, into time, into finitude: its felt irretrievable "past tenseness" is what we will mean by *suffering*.

Becoming conscious of our tail-chasing questions, a consequence of suffering experiences, we may throw up our hands and call such inquiry "absurd"[9] – or we occupy ourselves with myopic problem-solving, or become master *self*-story-tellers, or we desperately seek ways to regress to non-conscious states. Can we reverse evolution and stop thinking so hard and enjoy being alive *as we want to imagine* animals and small children do?

But there is no epistemically responsible escape from consciousness or its consequences. The very fact we perceive there is something to flee or dismiss is *already* a product of consciousness. Nostalgia for developmentally primeval states, entails our unhappiness with the one we are in. Happiness is a judgment, not a state.[10] If all that counts for happiness is *being* something or other, being in a state of non-pain: the ecstatic and the dead are happy because they *fully* are what they are. Their being is a unity. They are *not* observing themselves. The verdict that one is happy is fraught with an awareness of the alternative: *non*-happiness, something we must *also* be imagining if we are even considering the question. The faculty of *imagination* is a product of consciousness. To imagine is to picture counterfactuality. What would motivate such an effort and capacity if the coldest and hardest available facts were not sufficient for us? Nose to the ground, if the ground is adequate to your complete needs, why raise it? Whence this need for seeing beyond the view from the ground? You must have glimpsed something at the horizon not to be content with your present elevation.

It isn't clear whether consciousness came first, then its collision with the world it enabled access to caused suffering – or whether it was pain, or the evolved capacity and effort to avert it, that provoked the increasing awareness we call suffering. Perhaps consciousness and suffering emerged together, perhaps they are phenomenally *inseparable.* Something close to this is what we are suggesting.

You must be conscious to suffer and, if you are not suffering, you are not conscious. Indeed, you may be *fully* alive, even in pain, *non*-consciously, while *not* suffering. Two examples: being in *the throes of pain* which precludes the luxury of contemplating your condition; but just as well: *immersed in routine* – neither jarred by pain, nor particularly aware of anything. Reflective consciousness,[11] then, is of no benefit to you if you are *fully* alive, that is, engaged in the business of living, and not, as Socrates recommended, examining your life or evaluating whether it is worth living. The moment your aliveness becomes sensitive to its precarity,[12] suffering looms imminent. This sensitivity – the vulnerability of a living, normatively-equipped, state – again, is what we mean by suffering.

Pain has minimal requirements in the way of cognitive intervention. A nervous system will do. Evidence for pain's presence need not involve the complications of overwrought awareness. Pain-aversion is observable all the way down the ladder of life. Plants observably express their aversion or

9. Absurdity is a *concept* as self-defining as they come,

10. Suffering may be a judgment as well, but wallowing in illusions about it (as with judgments of well-being or happiness), has *less* evolutionary support. Not none, but less. The environmentally-conditioned pessimist, cynic, melancholic, or depressed have not been naturally selected out of existence. They survive recessively. Why? Occasionally consulting a "realist" (often another name for these) must also contribute to species survival.

11. Not to be confused with *access* consciousness, shared by most living things not stunned or asleep.

12. Consider, for example, the poem by e. e. cummings, "It may not always be so" and the music by the Bruno M. Miranda it inspired.

preference for certain environmental conditions. Pain is noted when an entity expresses through its behavior that it is having what, to it – all objective indications are – must be a negative experience. To be acknowledged as "pain," it must also be interpreted as such by an observer, *sufficiently like* the observed to imagine being in that state themselves. This sympathetic reaction also involves reflective consciousness. In other words, pain is not a concept, except to an observer. When such an observer is the same as the one who is having the experience, it is not only noticed (as pain is), but *experienced* as suffering.

For those actively in its throes, pain does not tolerate concepts – even and especially that of *time*. Suffering requires much more counterfactual ability: imagination, chief among them: a notion of perdurance across time, capacities to remember and anticipate, that is, backward and forward time-tracking.

Types of suffering

Not even the cat is seeing objectively, though it gets closer. There is less cognitive inflection in its judgment.

Some thinkers affirm a typology in the phenomenology of suffering.[13] One kind of suffering can be conducive to a development that some call "life-enriching," at least, if not quite "-affirming." *Mourning* and *regret* have been described this way. The idea is that *processing* these states will make you, in some sense, *stronger* (if they don't kill you first, as Nietzsche put it). These may be *construed* as "learning opportunities."

Melancholia and *remorse*, on the other hand, have been judged morbid variants of suffering. These may well kill you… or make you so vulnerable to consciousness-ending events that they may as well kill you *because they alter your capacity to value what you may have valued previously*. In any case, you never recover intact enough from these permanently scarring states to celebrate again with the same abandon. You are, in an important way, crippled[14] for life. If you survive, much luck – *keenly evident to you* – will be involved. You have been placed "on notice" and are not allowed to forget. Much for your success will depend on your capacity to re-describe self-deception as "self-remodeling." Failing that, this kind of suffering may attempt to embrace itself for what it is, and judge life as trauma waiting to happen... from which only animal forgetfulness may provide respite. Still, "may" is not "will." The *valence* of trauma, if not its phenomenology, may change (as we will see in the last section).

A "healthy" personality digests suffering, extracts what nutrients it may, and excretes the rest. But is *health* what it's all about? What for, *health?* To live and suffer another day? Is it supposed to be obvious that this is a *good* thing? The very concept of health is compromised – post such awareness.

To live a "full life" means what? To have your senses, sensibility, and understanding dulled into quiet acceptance of pending oblivion? To live a "full life" means to give up some experiences, those that would temper its "fullness," that would line it with ambiguity. Is there nothing to be gained by

13. The typology alluded to in this section, owing to Vladimir Jankélévitch, is discussed by Matthew Devine in "Suffering the Eternal: Remorse and Melancholia through the Work of Vladimir Jankélévitch," The specific typology strikes us as a bit arbitrary, though interesting.
14. Or we say "more mature," depending on how facilely you avail yourself of the psychological mechanisms Metzinger discusses.

forfeiting a subdued sanguinity? Is there some wisdom or understanding you miss out on to avoid this temperance? If some experiences may shorten your life, does that always count against them?

Judgment is expressed here. Normativity is in the air.. What do you *ultimately value*? Does *more* life take priority over anything learned by placing your priority elsewhere? Is there no "elsewhere" for you?

Conscious or not, suffering or not, all capable of it die. The suggestion here is not that you should change your *ultimate* values but that you should allow the existence of others than those you have. Fear of being alone may make this hard. Much thinking about these questions is commonly centered on the priority of prolonging life without further consideration.

...

It may seem, to the "healthy" majority (the ones that tend to our non-extinction), that *the conscious and suffering*, the melancholic, never having "lived fully," lose more by their intransigent will to awareness. Their one chance at happiness, fulfillment, transcendence, salvation, they forfeit. For what? What's to be gained from having one's eyes open overly wide? These and their epistemic curiosity, verging on vanity, will perish together. Maybe…

We grant this. After all, to those finding any measure of sweetness in it, melancholy is made sweeter in knowing it is always under siege by the forces of life. The arrogance, even the malice in this attitude, of pitting oneself against whatever bleakness is in store, notwithstanding.

Except that, *sometimes*, it seems the conscious and suffering positively *enjoy* their condition and would have it no other way – this by way of explanation for nature's tolerance for them. It is their proprietary bias. Even from a purely evolutionary perspective, there must be a reason, these depressed melancholic, party-poopers have not all been selected against long ago. Whether it is their intent or not, their persistence, even thriving in the hostile environment of first- or second-order hedonism, must be contributing something to everyone else's project, if not always their own.

Metzinger's conditions of suffering[15]

1. Consciousness (C-condition) – the experiencing entity must be basically aware. It must have *access* consciousness, that is, is capable of being environmentally-informed.[16]

2. Phenomenal self model (PSM-condition) – the experiencing entity must "own" its suffering. It must picture itself as capable of having experiences, of *possessing* them. (Historical, but less precise, expressions, used for expressing the self-modeling idea have been "subjects," "souls," "selves," and "minds.") Metzinger wants to say *there is a thing and it owns*, so to speak, states or experiences. Thus, there is an ontology of *experiences* and *their owners*.

3. Negative valence (NV-condition) – the experience must inspire aversion. To be suffering, the experiences of the owning thing are normatively devalued, *i.e.*, "bad." This implies the owning thing must be normatively-equipped.

15. See Metzinger, op. cit., p, 245, for his wording.
16. Know its milieu. Be existentially-informed. Behave like it, at least. Sometimes.

4. Transparency (T-condition) – the experience is "real," beyond subjective doubt, not dismissible as speculative; the connection of the owner to the experience is direct. It is not mediated by representation: I cannot think in the throes of the experience that my experience is the result of this or that perceptual or circumstantial or conceptual mediation. I don't need to ask you or consult instrumentation to be informed of the experience. I cannot channel Berkeley and think it is all in my mind as I cross the street in front of a speeding bus or contemplate the marvel of my nervous system as my tooth is being drilled. There is a clear unobstructed path or connection from the experience to its owner.[17]

The theory is that when something meets all these conditions it is suffering. Obviously, people suffer. Most higher animals do. How far down the ladder of life is an open question. What about an artificially and generally intelligent thing? We would need to check off items on this list to see.

The first three can be empirically corroborated in the case of animals or machines to the *same* extent as with people, using the same methods. These have behavioral markers. The corroboration involves their interpretation. We do this everyday when interacting with members of our species. I don't *know* that you are suffering except by *guessing* that you are. My guesses may be so powerfully corroborated that only a philosopher could raise questions. They are, nevertheless, guesses.

The transparency condition implicates the "other minds" problem. To the extent we surmount that problem among ourselves, we are not precluded from doing it for any life form or artificial entity behaving similarly.

If we don't extend our imagination to do so, we are under a moral obligation to explain *why*. And if we can't explain why, we have expanded the sphere of ethical concern.

Why is having a theory of suffering important?

Aside from the purported[18] value of Socratic human self-understanding and constraints on our treatment of animals… there are implications for AI: because it may tell us when we have succeeded at reaching a level of artificial intelligence that will make ethical demands on us. When an AI expresses the *necessary* conditions we need then to ask whether the *sufficiency* condition has not also been met. If it has, we need to worry.

To those who think we are still a long way off from creating the *artificial* suffering attending *artificial* consciousness, Metzinger cites the view of many eminent thinkers[19] at the time that it was "impossible" for something "heavier than air" to fly – even years after the flight of the Wright brothers.

So, Metzinger argues that we should consider *banning* AI research until we know better what will follow if we succeed.

17. As with Tamar Gendler's notion of an "alief," an emotionally-laden "belief." (Urgency, not to say, life thus distorts reality. If the dead could do science, they would be better at it than we.)
18. We acknowledge the possibility that Socrates may have been wrong about the unexamined life not being worth living. Maybe a not too fully unexamined life is the *only* one that permits of value. Too much examination is either distorted by motivation, in which case an affirmation might have been reached sooner less the energy expended on the effort of looking for it – or concludes in a demurral on the case for perdurance. Other outcomes are epistemically suspect. We, the conscious *and* living, are interested parties.
19. Notably, including mathematical physicist Lord Kelvin.

His qualm with AI research, to the extent the “intelligence” in “artificial intelligence” seeks to approach anything like what we mean by *conscious* intelligence, is that since it is *we* who define intelligence and consciousness, it is *we* who acquire the moral burden.

A light switch is minimally intelligent, but there is nothing even vaguely involving ethics in exploiting its intelligence.[20]

A laptop? Still not.

A computer *programmed* to optimize and preserve its structure or function? We are getting closer.

A computer *expressing*, through its behavior, optimization or preservation of its structure or function? Perhaps we are there. Decisions about why or why not we should consider its moral agency or patiency become relevant. A *system* – any kind – that behaves recognizably in this way – as people and all living things do – becomes thereby an object of moral consideration.

Shall we design our intelligent, potentially conscious, creations with biases to match our own? Shall we make them gluttons for punishment – all the while, happy enough to preserve their structure and function? Shall we divorce them from reality as we do ourselves? Isn’t there enough pain and suffering in the universe without our contributing to it? *Must* we follow in the footsteps of nature or alleged intelligent designers?[21]

Or must we regress toward *non-reflective consciousness, i.e., access* (only) consciousness? As much regarding ourselves as our most advanced creations?

Watch out when artificial general intelligence begins to show signs of giving a damn.[22]

The benevolent, omnicidal, antinatalist robot

Suppose an artificial general intelligence (AGI) is programmed with negative utilitarian normative firmware: pedestalizing the minimization of suffering: *a little* suffering is better than a lot, but, better still, if realizable, is *none at all*. Since, presumably, only living things suffer (so far), and only living things reproduce suffering by reproducing themselves, why not nip the proliferation of suffering in the bud by eliminating these suffering suffering-generators? That is, on *ethical* grounds, from sympathy with the idea that suffering is a bad thing, we trade a little suffering[23] now for preventing an inconceivably larger amount later. (This AGI would thereafter eliminate itself, too, of course – lest suffering, if it had not already, evolve again.) This AGI would be, *ex hypothesi*, *both* smarter *and* less selfish than we are...[24] This is one way out of the problem of suffering Metzinger considers.[25]

---

20. Leaving aside, for now, panpsychist understandings of consciousness. For now, these are not scientifically tractable.
21. The argument is akin to that of some Benatarian antinatalists or nonexistentialists. Metzinger notes the resonance and extends ethical concern to artificial forms of suffering that may attend developments in AGI – especially, if we don’t have a clear understanding of how a self-modeling, reflectively conscious, entity may emerge from our efforts.
22. See our discussion “Robot Ethics: Part 2 - Does “no” mean “no” when a robot says it?”
23. The only suffering imposed by euthanizing us would be the frustration of our future expectations – an objectively small price to pay for the glory of having produced a modest region of the universe free of suffering.
24. What would be the point of making robots as dumb or as selfish as we? Or worse, more of both?
25. See Metzinger, op. cit., p. 253.

Let's spell this out more. It is important to specify our terms. Not to do so is to invite arbitrary conclusions. Arbitrariness is irrational and immoral if anything is. Our artificial system would be designed for optimizing rationality and morality. To be rational means to work out a conclusion from explicit premises employing only perspicuous rules without misstep. Everything must be clear and explicit: the premises, the rules, and the conclusion. The cardinal moral premise here will be simple: *pain and suffering are bad.* Badness, to the extent possible, *ought* to be eliminated. Our system will be supremely intelligent, capable, and rational. In addition, we want it to be *compassionate,* to be averse to the presence of preventable negatively valenced states, e.g., pain and suffering. To be compassionate is what it is to instantiate morality.[26] For this system, being rational and being moral will be imperatives. These are its most fundamental principles. It takes its imperatives *dead* seriously. Unlike human beings, as we know them, who are pushed and pulled in many seldomly fully coherent directions at once, our artificial system will not deviate from its marching orders. To the immediate objection that it cannot be compassionate to euthanize extant life against its will because of the suffering this will cause – that is, because it will painfully frustrate *this* extant life's will to live, the system will answer:

> It *is* rationally and morally justifiable to do so because of the vastly *more* suffering it will forestall. Not to do so would evince a selfish preference for the *proximate*, for preventing a *smaller* quantity of suffering over an unimaginably *greater* amount even if more *distal*.

Nothing in this artificial system's premises indicates that it is rational or moral to *overvalue* the immediate over the inevitable. Were there, it would go against one of the most widely recognized moral principles: *impartiality: I cannot say that my pain is more significant than yours.*[27] Space, time, or affinity does not change this. It would enshrine rational and moral arbitrariness, otherwise.

...

By contentedly keeping the concept of what constitutes consciousness vague, by not specifying scientifically tractable markers of it, *such as suffering*, we excuse ourselves from thinking too clearly about what might happen if we press on with developing systems that may approach it. Thus, a call for a moratorium on AI development by many AI theoreticians and researchers until we know better what we are doing.[28]

Moral principles at stake in Metzinger's thought experiment

Three widespread moral intuitions together lead to the alarm suggested in Metzinger's "benevolent, omnicidal, antinatalist robot" thought experiment.

---

26. The negative utilitarian premise, of course, is controversial among philosophers, but almost everyone can accept it *prima facie* – though it seems to me only by proceeding to surreptitiously subvert it altogether with qualms and distinctions, as did J. S. Mill when he inserted qualitative hierarchies of pleasure and pain. A different argument to the same antinatalist conclusion founded on Kantian principles was made out by Otto Weininger in 1903.

27. Consider ditching this principle, Aristotle-like. Make all ethics local. The slope from this to increasingly myopic self-interest is steep. Concern shall be restricted to myself and those with whom I assert affinity. It will be pointed out that no ethical system escapes this tendency, not sentience-based or reason-based theories, since these each assert minimum qualifications for entry into the circle of concern... Perhaps, ethics is incompatible with the existence of *conscious* systems: the access sort because it is not sufficient to formulate one; the reflective sort because it will mandate itself out of – or excuse itself from – existence. Weininger concluded as much...

28. Metzinger asks for a pause of 50 years. The signatory list for the moratorium asks for a more modest six months.

1. **It is *more* wrong to bring about pain and suffering than it is to *fail* to bring about pleasure and happiness. (Benatar's asymmetry.)**

This idea is pervasive in ethical and legal thinking. It undergirds, for example, such phrases as:

- "If you can't help, at least, don't make matters worse,"[29] "if you can't be part of the solution, don't be part of the problem," or "we should leave well enough alone."
- "First do no harm."[30]
- An accused is innocent until proven guilty.[31]

These moral rules of thumb or heuristics adjure us not to take unwarranted risks that may bring about worse outcomes in order to promote the possibility of good ones. The good ones *ought* to be overwhelmingly probable to override the conservative principle. Among the fears are that we may *well-meaningly* kill the patient or punish the blameless.

We see the principle at work in practices such as:

- We don't criminalize people who *fail* to give to charity.
- We don't punish people who *refuse* to risk their lives to save others.

There are exceptions to these rules but they invite controversy:

- A pregnant woman who chooses to abort over proceeding with a birthing that will risk her life.
- A person who refuses to take up arms to defend their country.

In such cases, there either are, or have been, moral sanctions, if not also legal ones, against such behavior, *as though* there are some things for which we *ought* to take extreme risks because of a purportedly great good that may come of them – one judged worth risking a very real possibility of a negative outcome.

But, by and large, we think *worse* of people who *deliberately and recklessly* bring about pain and suffering even though they and others may see themselves as do-gooders[32] than we do of billionaires who *could* relieve the suffering of hundreds of thousands, at small cost to themselves, but *don't*. Why, exactly, we do this is an interesting question. Some philosophers think we should not behave this way: we ought always to place *at least* equal, if not more, emphasis on promoting good outcomes than on worrying over the possibility bad ones.[33] (Positive or neutral utilitarians, for example.) We won't pursue the priority question here. We only observe that the conservative principle points to an asymmetry in how many of us, in fact, weigh good and bad experiences.[34] We seem to think that *it is more urgent not to do wrong than it is to do right.* That seems the default position. Mostly, our institutions reflect this intuition. When they don't, eyebrows are raised. A case has to be made for going against the intuition.

29. A version of this is attributed to the Dalai Lama.
30. See this more sophisticated discussion of both the impracticality and *ultimate* ethical import of the abbreviated phrase.
31. Found in the Code of Hammurabi 18th century BC.
32. What is judged as "reckless" may require a high degree of critical mindedness to see. *Too much* deference to obviously compromised authority, as in the recent pandemic, is a case in point.
33. Certain communitarian and Marxist intuitions are suggested.
34. Again, as famously remarked by David Benatar.

Thus, when doing something may have good consequences (such as promoting pleasure and happiness) but also may have bad consequences (such as causing pain and suffering), we ought not risk it unless we are quite sure the good will outweigh the bad. Indeed, implied also is an imperative to *determine* the risk accurately. *But there are cases where this assurance can never be had.* One such is the production of entities capable of both positive and negative experiences, where the negative ones are *assured* while the positive ones are not. Being born is an example.

Suppose, unlike the typical case, a sentient being is born into the most fortunate circumstances imaginable. Everything goes right from birth to death for this being. Its life is replete with *only* positively-valenced events. Assuming only unrestricted, unalloyed joy and happiness are good things,[35] then it seems the ending of its blissful life would be *especially* tragic – more so than for a life already inured to coping with some bad experiences. More is stood to lose for the blissful life. Assuming the entity appreciating its own existence is not immortal, it will one day have to face the end of its existence: it will die. How else is it to face its demise except negatively? Again, unless it develops a taste for some negatively-valenced experiences. And how could that happen?

Might there be a consolation offered by a "merciful" evolutionary process? Might the hapless existent learn to adjust what it values in order to face successfully its current, or impending, reality? ("Successfully" meaning here: not giving up, not courting death or extinction, prematurely.) Adjusting one's values to fit the situation is also known as a bias: a self-addressed narrative, disconnected from a vision of "objective" reality, devised to make one feel better about things… which would be good for survival (as Metzinger notes) *were there recognizable continuity between what was before and what survives the adjustment.*[36] Shall we say that the thing that was born is *not* the thing which dies?

The question becomes sharper if the entity is presumed immortal, or may become so: to what extent are we – *equipped* as we are, *evolved* as we have – to identify with such an entity? Our values are what they are because they were conditioned to be that way by every experience we and our evolutionary precursors have ever had. There are real limits to the scope of what we may imagine – barring transformation into something we are *not* now. Becoming *something else*, then, why would we be concerned? My body (presumably, also my "mind") will one day degrade into the elements on the periodic table from which it is composed. Can we really imagine identifying with the motley collection of those elements and the forces that act on them sufficiently to care about their fate across the time and space of an, as yet, non-existent universe? It strains altruistic resources to care.

2. The elimination of pain and suffering, to the fullest extent possible, *is* a moral imperative. (A fallout of utilitarianism.)

Not everyone measures good in terms of non-negatively-valenced moments. But even animals and children seem to. This is a reason why, among ethical theories, some form of utilitarianism is often the first theory introduced to moral philosophy students. Pain and pleasure are so easy to grasp. At least initially. Don't cause the first and do the second. More sophisticated theories such as Kantianism presume so much that is not naturalistic. Science is, it seems, more comfortable with phenomena

35. And not more complicated or nuanced valences as discussed earlier.
36. The non-identity problem: altering characteristics defining of an entity across time makes identity implausible. Is mortality defining of living things?

amenable to empirical observation… so let's go with this *prima facie* sensible rule. Define which phenomena are negative and which are positive and even a machine can be programmed to minimize the first and maximize the second. We may devise a thing whose behavior we may have clear criteria for evaluating. This is as simple and sensible as moral theory ever gets.[37]

3. **We are under an obligation to *improve* on our failings. Morality is progressive. There is no such thing as a level of moral development with which we may rest content. (A perfectionist principle: evident in varieties of virtue theory and Kantianism.)**

These ideas combined seem to entail that *if* one effective path to our moral improvement leads to creating artificial entities *more* capable, *more* rational, and *more* benevolent than ourselves, then we *ought* to create them. Indeed, we *ought* to create the conditions fostering their creation. We ought to make artificial agents *morally superior* to us. Our own standards of what would make for a better world demand this. This is a moral imperative.

There is also an inductive argument for the conclusion that this *will* happen quite apart from whether it *should*. As a biologically evolved species, we are indeed the most capable of those we know. But we are beginning to envision entities more capable than we. If we may do more than merely envision them, if we are, or soon will become, equipped to bring them about, we *will* do so. We have no history of not doing what we are capable of. If we see any near-term advantage in doing so (and it is obvious that we have and will), we will push our technical capabilities to the limit – irrespective of the morality, irrespective of any radical or far reaching consequences. This appears to be evolutionarily determined. Nature is *not* concerned with morality. Nature is *not* concerned with "forever." *We are processes that play out.* At best, we observe nature expressing a preference for variegation over long (to us) stretches of time. But *never* forever. We and our normative baggage born of sentience, too, will pass.

Thus, we *must* strive to make things better. It is incumbent upon us because of the normativity of the third principle. It dooms us – to the project of creating an entity *more* capable, *more* intelligent and *more* benevolent than we can ever bring ourselves to be. It would kindly put us out of our misery.

It *would* be a coup de grace if we *willingly* brought about our own extinction because we had come to see it as the right thing to do: to go out willingly and gracefully. It would be aesthetically sublime if it turns out that way. Not as unknowing victims as with the dinosaurs, not kicking and screaming as with perhaps so many hapless species before us. We may yet rise above our animality and immaturity…

But also – we may never. We have little history – biological or cultural – of instantiating gracefulness. All this, even assuming it is up to us.

...

We *would*, either way, have participated in a grand piece of installation art – *were* there non-culpable witnesses. But that would be a contradiction. The museum the universe is and always was will again have to remain unvisited until some random time, if ever, the production of witnesses started up again. The production of art ready for appreciation as such would suffer a hiatus.

37. As moral theories go, utilitarianism may describe less accurately human normativity than, say, Kantianism, but it has the virtue of being easier to conceive and implement.

…

We might, of course, discard any one or all of these three moral intuitions. Where would that leave us?

Defining our way out of the dilemma

Why don't we just *stipulate* that non-biological entities can never be objects of moral concern by defining sentience, normativity and any consciousness arising from them as possible *exclusively* in *biological* contexts? If it ain't alive, we needn't worry. If it ain't alive, nothing of moral concern can arise about how we treat it. AGI, no matter how capable, intelligent, or benevolently programmed and behaved, will always remain only of instrumental value *to us*. It will only ever have the moral status of an eating utensil.

Three things:

First, mere utensils have been used by us to kill ourselves, no matter their moral status. Electricity, chairs, ropes, sharp objects, toxic and explosive chemicals etc. are, and have been, quite useful to our survival. Metzinger's "benevolent, omnicidal, antinatalist robot" – call it "utensil" if you like – would still serve "our better angels" even as it exterminated us. Recall: the values it would execute will have originated *with us*. And to the extent we identify with those angels, we may wish it luck.

Second, if only it were true that non-living things or inanimate objects were never objects of moral concern. But they are and have been. Consider the fierce debates in environmental ethics as to whether we *only* care about the environment because of its effects on the quality of life we expect and want our descendants to have: this is an anthropocentric argument. It is just selfishness writ large.

Or is the environment valuable irrespective of any interests we have? Aren't landscapes, ecosystems, sunsets, mountains, rocks and even lifeless planets beautiful and worthy of care and moral regard no matter whether we or anything like us is there to appreciate them? Some people think it is good that, as yet, there are no McDonald's burger wrappers on the moon. Was the earth worthless before we came along? Millions of years of evolution significant only because it led to us? Was all creation of value *only* because we appeared? Biblical pronouncements aside, many can imagine these are objects of moral concern, that we ought to tread lightly and not just because it may matter to whatever results from the survival of our kind.

Third, whether or not any entity has or hasn't any value, apart from its service to our interests, is *not* entirely a function of any "objective" property it may have. Beings like us *confer* value on things that are not and never have been alive, and not just because of their usefulness to us. Imagine barbecuing a Teddy Bear, drowning a Barbie doll, or watching a hot-wired old clothes dryer rip itself to pieces trying to "dry" a piece of iron in someone's backyard for the amusement of Youtube viewers[38]… Psychologically, we evolved to have our feelings triggered by the "apparent" suffering of *anything*, dead or alive, that even remotely resembles something we associate with suffering. It doesn't even have to look and behave much us like us: it is enough if it does one of those things in certain contexts. AGI

38. Even as a wildly popular spectacle, some viewers of the "dryer kill" found it "disturbing," even "terrifying" to watch. Why?

robots will pull our “heart strings,” if we have any, the more capable they become of constructively *interacting* with us.

When AGI robots begin to trigger our feelings as they already have, and inevitably will more, we will begin to insist that our circle of moral concern expand to include them. Moral relevance may lie in the spaces between apparently conscious entities – in their interaction, their relationships – as much in any ontological features properly theirs. Morality, like beauty, may be in the eye of the beholder.[39]

Bittersweetness, love/hate, aficionados of aversion, tragicomedy, and paradox-mongering…

There’s another aspect of suffering that is seldom remarked in analytic/scientific discussions. It probably doesn’t apply as much, or at all, to pain, but the negative valence paradigmatically attributed to suffering is not always unadulterated. Suffering can be pleasurable, an object of pursuit. It can be something valued in and of itself, and *not* just instrumentally – because it may be required to attain something less ambiguously so, such as the reward or recognition that comes of hard work. Sometimes scary or aversive things attract us – and *not* because we are unaware of the pain, danger, or mortal or morbid consequences attending them.

People jump out of planes for fun. They fall in love, not always ignorant of the real probability that it may not go well. They develop a taste for the bitter, the darkness, the disappointment, and the pain preceding and at the moment of death. Children find their first taste of alcohol or vegetables unpleasant, not like cookies or ice cream. Still, for better or worse, many come to find these initially unappealing experiences addictive. Health alone doesn’t always explain healthy life styles. Some people exercise vigorously for the shear altered state of it, pain and all. Masochism is real. We develop a taste for it. The chance to contend with suffering may become a second nature skill we would miss not having. Novelist and Nobel Laureate Isaac Bashevis Singer was once asked, long before it became fashionable, why he was a vegetarian. We paraphrase his answer: “For health reasons... Not mine, the animal’s.” It’s not just that sometimes we value the consequences or want to live longer. Sometimes we suffer for abstractions or ideals that few, pretending to rationality, will understand, let alone value. Just as Kant was willing to live with the logical consequence of his moral theory: that, *as far we can know*, not once in the history of the species has a single moral act ever occurred. Notwithstanding such dismal encouragement, Kant was not alone in almost reveling in the absurdity and paradox of the imperatives he thought harassed us. He knew exactly what his moral thinking entailed. It is why the moral law within was as inscrutably fascinating as the starry heavens above.[40]

I can’t go on, I will go on.

– Samuel Beckett, *The Unnameable*

---

39. Kate Darling, an expert on the moral and legal implications of technology, explores the idea that the capabilities we may attribute to “robots” – sentience, consciousness, and moral status, among them – have as much to do with us as with the target of the attribution. Modern technology may enable a better *subjective* understanding of the prehistoric notion of animism. We are learning to feel and find significant what, in one sense, is not there – *again*. See her TED talk and interview: “We shouldn’t laugh at people who fall in love with a machine. It’s going to be all of us.”

40. “Two things fill the mind with ever-increasing wonder and awe, the more often and the more intensely the mind of thought is drawn to them: the starry heavens above me and the moral law within me.” – Kant’s famous line from the conclusion of the *Critique of Practical Reason*.

People who suffer are not always innocent. We sometimes court disappointment fully cognizant of what we are doing. Some criminals, epistemic as material, have no need for the spoils of crime, the thrill of getting away with it can only be a thrill if they may *not* get away with it. This can be enough to motivate.

...

We suffer and – sometimes – would have it no other way.

If love's a sweet passion why does it torment?
If a bitter, oh tell me, whence comes my content?
Since I suffer with pleasure, why should I complain,
or grieve at my fate, when I know it's in vain?
Yet so pleasing the pain is so soft as the dart,
That at once it both wounds me and tickles my heart.

– Henry Purcell, *If love's a sweet passion*[41]

Nietzsche called it *amor fati.*[42] It is a kind of transcendence of hedonism. Whatever comes, up to and including death, I am game.

…

Again, it must have been useful, *some* of the time, for *some* of us, a minority, to be sure, to develop a knack for survival in the full face of the pointlessness of doing so. Not from sanguinity – for this minority, at least – but from realizing the equal pointlessness of not surviving. How useful? Because sometimes – if not most of the time, admittedly – a witness to the outer reaches of conscious awareness may conduce to the momentary, aesthetic experience of seeing things for what they are: to see experience and its self-modeling witnesses as bewildering contradictions. Again, to the extent this is ever "seen" at all.

Consciousness arose from a need to step outside mere animality or the environmentally predetermined. Consciousness may be judged an illness to the extent it cripples sanguinity.[43] It may not be life-affirming. Having stepped out, new reasons to persist are demanded. Even a willingness to step out of consciousness. *If* persistence, in recognizable form, is still in the cards.

Does intelligence imply rationality?

Near the start of this writeup we suggested that *if* our survival depended on happiness and *if* we were rational, we should by now have given up. Perhaps there is a mistake in assuming, as it seems we are

41. From the *The Fairy-Queen* (1692). Three versions of Purcell's air: Elizabeth CD Brown, Kronthaler, and Les Arts Florissants.
42. In "Why I Am So Clever" *Ecce Homo*, section 10, he writes: "My formula for greatness in a human being is *amor fati*: that one wants nothing to be different, not forward, not backward, not in all eternity. Not merely bear what is necessary, still less conceal it—all idealism is mendacity in the face of what is necessary—but *love* it." From *The Gay Science*: "Only great pain, the long, slow pain that takes its time... compels us to descend to our ultimate depths... I doubt that such pain makes us 'better'; but I know it makes us more profound... In the end, lest what is most important remain unsaid: from such abysses, from such severe sickness, one returns newborn, having shed one's skin... with merrier senses, with a second dangerous innocence in joy, more childlike and yet a hundred times subtler than one has ever been before."
43. Wittgenstein alluded to the need to philosophize in pathological terms.

here, that when we talk about intelligence we must be talking about rationality, too. If that's the case, we get the alarming result of Metzinger's thought experiment. One way to avoid the alarm is to consider that rationality is *not* a necessary, let alone sufficient, condition for intelligence. Suppose *useful cleverness* is a better understanding of "intelligence" than anything to do with rationality. The strict boundaries that rationality places on thought and behavior would be dropped. *Any* thought or behavior that served the current ends of a self-preserving system would count as "intelligence." This means that a lot of unabashed irrational thought and behavior, depending on whatever arbitrary ends the system had, would be acceptable. This describes the behavior of humans and other living things just as nature made them. Reason is a tool, *only* a tool, and can never set agendas. When life squares off against reason, we know which has won – *so far*. We are living proof of this. Surviving systems operate according to more or less stable whims. For finite periods. They may pursue ends that contradict and/or sabotage other ends they simultaneously pursue. We might wish to insist that this *still* manifests intelligence.

If so, and we want to toy with creating artificial systems that mimic or instantiate intelligence so described, we will need to program these systems with competing, even self-defeating, algorithms. Under such a rubric, a spell checker will sometimes call-out correctly spelled words as misspelled, it might reference a library of mutually inconsistent dictionaries, and randomly defer to one on any given occasion. This is how living things sometimes manifest intelligence. It may not be rational but still intelligent. The criterion is: does it work to achieve the particular end in focus? The end might be poetry. This is compatible with the left hand ignoring what the right is doing. Morally, it will be perfectly OK to rob Peter to pay Paul. Paul may be present, Peter not. Or the reverse. In short, AGI will have to think and do stupid things to mimic human-like intelligence. But what will count as "stupid" will require revision...

Perhaps we still don't know what we are after in pursuing the creation of artificially intelligent systems. Do we really want to copy something like human and animal intelligence?[44] Or do we want something that manifests all that we have, at times,[45] considered *best* about ourselves, not what we are as nature made us? Something, in everyway possible, *better* than us? At times, we have thought that being rational and moral were desirable behaviors. Maybe we don't anymore. Drop these qualms and the path forward is clear: we will do what are going to do until something external to us or some whim of our own stops us. Hold on to our rational and moral squeamishness, and maybe *we* will author our final chapter.

Alternatively, AGI may be designed to preserve a normative status quo. We don't want to get better. We want more of the same. If only we had any reason to believe that even this was a stable whim.

The *discovery* of suffering through reflective consciousness

Pain was there from the beginning. Suffering was *discovered*... The way out of the non-zero probability of facing the benevolent, omnicidal, antinatalist robot dilemma is to remark the *mistaken* premise that

44. With all the beauty and horror it is capable of.
45. We are still living in the shadow of the Enlightenment when such sky-scraping ideals were articulated and celebrated. The premonition that our ideals may leave us "happy in the lot no created thing" is described by the Renaissance philosopher Giovanni Pico della Mirandola, and quoted by Otto Weininger in his infamously antinatalist book *Sex and Character* at the end of Part II, Chapter 8.

pain and suffering are bad. That it is error to think any truly moral principle would call for their elimination. “Made the way we is,”[46] that’s how we go.

Resources

“Suffering,” Thomas Metzinger, in *The Return of Consciousness: A new science on old questions*, Kurt Almqvist & Anders Haag, eds. 2017, p, 221. Highly recommended is Dr Tevin Naidu’s extended Youtube interview with Metzinger, “What is Pure Consciousness? Minimal Phenomenal Selfhood & Epistemic Agent Model | Thomas Metzinger.”

“Phenomenological Approaches to Self-Consciousness,” Shaun Gallagher, *SEP*, 2019.

Notes from Metzinger’s article, “Suffering”

> 240
> “Would you like to relive the very last conscious moment in *this* life?” Interestingly, under this condition only a little over 28 per cent of life moments were ranked as positive, while just below 72 per cent were considered not worth reliving....
>
> Of course, there are a host of justified, and important, theoretical questions about this thought experiment. Could I also take only the one very best moment of my life and repeat this single moment indefinitely? *When* do I have to make the final decision? Many positive experiences include an aspect of “novelty” or “surprise,” would this aspect be in definitely present in the afterlife as well?…

If the happy moments in our lives involve surprise, and surprise is only “surprise” against a background of expectation, does this mean “happy” moments require a background expectation of unhappiness or, at least, moments not worth repeating? If happiness is only noticeable against expectations of unhappiness or, at least, non-happiness, we must either not require much in the way of happiness to persist with our lives, or have psychological mechanisms for screening experience from an awareness of its lack.

> What are the temporal boundaries of “a” conscious moment? How does one introspectively individuate such allegedly “single” moments, and how theory-contaminated are the corresponding verbal reports? How *reliable* is introspective knowledge in the first place? And should we trust our own intuitive value judgments; can they be epistemically justified? And so on.
>
> The results were however striking: first, the initial surprise the low scores caused in all participants and, secondly, what could be interpreted as a consequent attempt to explain the phenomenon away.

Consider (1) the surprise at the low scores and (2) the “explaining away.”

46. Adapted from the line “made the way you is” from the story “Melanctha” by Gertrude Stein.

Indeed, “explaining away” a prediction error, in this case, may simply consist in an updating of one’s conscious self-model, preferably in a way that allows one to reduce introspective uncertainty by dampening and suppressing future “unexpected news” of this type. Many participants immediately [241] suffering started to develop more or less intricate “cognitive confabulations”, trying to make their own self-assessment appear not so bad after all: “Happiness certainly is not the most important thing in life, I am writing a doctoral dissertation that will make a contribution to the knowledge of mankind and epistemic progress certainly adds more value and meaning to my life than momentary hedonic valences or even their life-time average!”; “Most of my moments are neither good nor bad anyway, they are just neutral!”; “Life-time average value is philosophically irrelevant, it is only the peak experiences and our memory of them which make a life a *good* life!”; “What *really* counts in life are the larger time-windows, not decontextualised phenomenal snapshots!” And so on.

Narrative self-deception

If, on the finest introspective level of phenomenological granularity that is functionally available to it, a self-conscious system would discover too many negatively valenced moments, then this discovery might paralyse it and prevent it from procreating. If the human organism would not repeat most individual conscious moments if it had any choice, then the logic of psychological evolution mandates concealment of the fact from the self-modelling system caught on the hedonic treadmill. It would be an advantage if insights into the deep structure of its own mind, insights of the type just sketched, were not reflected in its conscious self-model too strongly, and if it suffered from a robust version of optimism bias.

Perhaps it is exactly the main function of the human self-model’s higher levels to drive the organism continuously forward, to generate a functionally adequate form of self-deception glossing over everyday life’s ugly details by developing a grandiose and unrealistically optimistic inner story – a “narrative self-model” with which we can identify?

Any successful agent must be able to motivate itself. One solution to this problem of autonomous self-motivation could have consisted in what I would like to call “autobiographical Gestalt formation”, an automatic escape into larger timescales. Maybe evolution has created self-models that automatically expand their predictive horizons as soon as the present is boring or simply too unpleasant? Such a strategy of flexible, dynamic self-representation across a hierarchy of timescales could have a causal effect in continuously remotivating the self-conscious organism, systematically distracting it from the potential insight that the [242] life of an anti-entropic system is one big uphill battle, a strenuous affair with minimal prospect of enduring success. Let us call this speculative hypothesis “narrative self-deception”. If something like this is true, one would also expect it to have an observable effect in academic philosophy and science as well.

243
...can there be conscious intelligence without existence bias?

244

Our constant inner battle for mortality denial is, however, only one introductory example. When one examines the ongoing phenomenology of biological systems on our planet, one finds that the varieties of conscious suffering are at least as dominant as, say, the phenomenology of colour vision or the capacity for conscious thought. The ability to see colour consciously appeared only very recently in evolutionary terms, and the ability consciously to think abstract thoughts, of a complex and ordered character, arose only with the advent of human beings. Pain, panic, jealousy, despair and the fear of dying, however, appeared millions of years earlier and in a much greater number of species. Now, as the environment human beings have created for themselves gets increasingly complex, the sheer quantity of preferences and thereby of potentially frustrated preferences continues to rise.

247
Phenomenal transparency means that something particular is not accessible to subjective experience, namely the *representational character* of [248]suffering the contents of conscious experience. This refers to all sensory modalities and to our integrated phenomenal model of the world as a whole – but also to large parts of our self-model. The instruments of representation themselves cannot be represented as such anymore, and hence the system making the experience, by conceptual necessity, is entangled in an illusion of epistemic immediacy, a naive form of realism. This happens, because, necessarily, it now has to experience itself as being in direct contact with the current contents of its own consciousness. What precisely is it that the system cannot experience? What is inaccessible to conscious experience is the simple fact of this experience taking place in a *medium.* Therefore, transparency of phenomenal content leads to a further characteristic of conscious experience, namely the subjective impression of immediacy.

This may also provide us with a deeper understanding of what the process of conscious experience actually is. To be conscious means to operate under a unified mental ontology, constituted by an integrated set of assumptions about what kind of entities *really exist.* Systems operating under a single transparent world-model for the first time live in a reality which, for them, cannot be transcended. On a functional level they become *realists*. Again, this does not mean that they have to possess or even be able to form certain beliefs, or use explicit symbol structures in communication. It only means that the implicit assumption of the actual presence of a world becomes causally effective, because, as philosophers might say, non-intentional properties of their own internal representations are not introspectively accessible to them – they necessarily experience themselves as being in direct contact with their content. This is also true of the conscious self-model. A transparent self-model adds a new metaphysical primitive, a new kind of entity to the system's ontology: the "self ".

[On AI suffering]

248
The phenomenology of suffering has many different facets. To give another example, artificial suffering in conscious machines could be very different from human suffering. It is also conceivable that, say, some kind of Bostromian "superintelligence" could represent negative expected utilities and frustrated preferences in inner forms of phenomenality that involve no conscious suffering at all. There could be perfectly rational artificial agents, exhibiting neither the biologically grounded "existence bias" nor any other of the human cognitive biases which result from the millions of years in which evolution has shaped the self-models of our ancestors.

But *if* they suffered, damage to their physical hardware could be represented in internal data formats completely alien to human brains – for example, generating a subjectively experienced, qualitative profile for embodied pain states which biological systems like ours could not emulate or even vaguely imagine. The phenomenal character going along with high-level cognition might transcend human capacities for empathy and understanding, such as through intellectual insight into the frustration of one's own preferences or into disrespect of one's creators – perhaps even into the absurdity of one's own existence as a self-conscious machine, a mere commodity or research tool used by an ethically inferior biosystem.

249-50
Of course, all four conditions specified here are necessary, but to understand the very specific phenomenology expressed by self-reports such as "I am certain that I do exist and I am identical with *this!*", the conjunction of the PSM-condition and the T-condition is central. The transparent world-model allows a system to treat information as *factual* information, as irrevocably stemming from the real world; a transparent self-model creates the Cartesian phenomenology of being certain of one's own existence. Any robot operating under a phenomenally transparent body-model would, phenomenologically, *identify* with the content of this model and hence with any negatively valenced state that might become integrated into it.

250
At this stage, our working concept of suffering is constituted by four necessary building-blocks: the C-condition, the PSM-condition, the NV- condition, and the T-condition. Given our current situation of epistemic indeterminacy, any system that satisfies all of these conceptual constraints should be treated as an object of ethical consideration, simply because we do not know if, taken together, they might already constitute the necessary *and sufficient* set of conditions. By definition, any system, whether biological, artificial or post-biotic, not fulfilling at least one of these necessary conditions is not able to suffer. Here the central desideratum for future research is to develop this first working concept into a more comprehensive, empirically testable *theory* of suffering...

We want it to yield *hardware-independent demarcation criteria* **[for suffering]**.

[Metzinger's negative utilitarianism]

...it is much more important to reduce suffering than to maximise happiness, because frustrated preferences are ethically more relevant than satisfied preferences. "Negative utilitarianism" (NU)

...almost nobody likes the idea of future artificial intelligences silently preparing for the extermination of all human beings and other sentient creatures on the planet – on purely ethical grounds…

254-5
...the phenomenology of suffering is not a simple mirror-image of happiness, mainly because it involves a much higher urgency of change. In most forms of happiness this centrally relevant

subjective quality which I have termed the “urgency of change” is absent, because they do not include any strong preference for being even *more* happy. In fact, a lot of what we describe as “happiness” may turn out to be a relief from the urgency of change. The subjective sense of urgency, in combination with the phenomenal quality of losing control and coherence of the phenomenal self, is what makes conscious suffering a very distinct class of states, not just the negative version of happiness. This subjective quality of urgency is also reflected in our widespread moral intuition that, in an ethical sense, it is much more urgent to help a suffering person than to make a happy person even happier.

[An intuition also defended by Benatar, for instance.]

255
One can also state and dissolve another frequent misunderstanding. Human suffering is rarely dramatic suffering. Almost all negatively valenced states involve only a mild emotional sense of preference frustration – perhaps some weak impairment of bodily wellbeing or a diffuse background feeling of boredom, possibly accompanied by an unspecific, generalised worry about the future plus a subtle phenomenology of uncertainty. In addition, as it is plausible to assume that these frequent and much more subtly negative states are forming the majority of our conscious self-model moments, most of us may have long ago begun to perceive them as inescapable and uncontrollable. We may therefore operate under a “domain-general” version of learned helplessness with regard to our own suffering: we become unable or unwilling to avoid subsequent encounters with such inner situations, because on a deep functional level we already believe that we cannot effectively control the total probability of their occurrence.
Consequently, we do not take action to avoid more subtle forms of negative everyday phenomenology.

Therefore, what we prematurely report as “neutral” states may often actually be the inner experience of subtle preference frustration plus learned helplessness. We report “neutral”, but what we actually mean is “default”. If one introspects carefully, truly neutral moments are some thing very rare, because some sort of affective valence accompanies almost all of our conscious moments...

the Philosophy club

Extended writeup for the topic hosted at

The Philosophy Club in October 2023

– Victor Muñoz

Guanajuato / Seattle